Progress Report for NASA Intelligent Systems Program (May 13, 2002)

Intelligent Engineering Time-Series Pattern Matching

- PI:
  - Dr. Dennis DeCoste, JPL
- Co-I:
  - Dr. Padhraic Smyth, University of California at Irvine (UCI)

Goal and Technical Objectives

- context: NASA missions/testbeds generate <u>massive</u> volumes of engineering time-series data but largely fail to exploit them
  - typically: millions of time points per week, thousands of sensors
  - largely checked in real-time and then *ignored* in future operations
  - ability to find similar historic data to current state (query) would:
    - help understand scientific or engineering phenomena (I.e. better designs)
      - e.g. find "thermal snap" events in SIM structures similar to one of interest
    - reduce cost of ops
      - e.g. find similarities (and associated corrective action logs) in previous Space Shuttle mission to currently detected abnormality
      - improve analysis/safety
      - provide robust basis for detecting abnormalities or known dangerous events
- **<u>our goal</u>**: *develop a fast search engine for time-series data relevant to given queries, suitable for real-time and off-line mission contexts (e.g. "Google for time-series data").*

Technical Problem Statement

- technical problem: find task-useful notion of “similarity” and fast way to apply it (i.e. avoid touching entire database)

- many technical challenges, including:
  - suitable “similarity” score is often domain & query dependent
  - traditional indexing methods quickly degrade to “linear scan” once dimensionality grows beyond 10.
  - thus, most other related research on “similarity search” assume similarity score function is given and focus on pre-query dimensionality reduction (e.g. PCA or FastMap), to enable fast off-line nearest-neighbor indexing methods (e.g. kd-trees, vp-trees, etc.).
  - however, time-series often impractical to reduce to ~10 predefined dims
    - multi-variant (many sensors), rich feature space (e.g. lags, frequency-domain), rich invariance space (e.g. scaling, shifting, time-warping, …)

Technical Approach

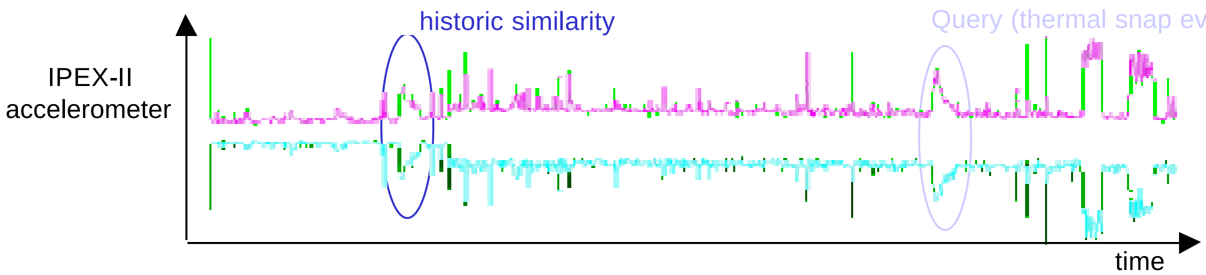
- our solution extends/combines multiple key ideas:
  - employ rich full set of possible time-series features
    - time lags, windowed stats (e.g. mean, max high-water marks), etc.
    - efficient lazy generation: only compute specific features from large candidate space as required during model training / selection search
    - dimensionality reduction innovations
    - support <u>nonlinear</u> reduction via kernelized FastMap & locally-linear embedding (e.g. [DeCoste, ICONIPS-2001])
    - new approximate nearest-neighbor (NN) similarity methods that exploit these nonlinear embeddings and reason about induced errors
    - but, most importantly and uniquely, focuses on query-relevance, via learning/exploiting *query models* …

Technical Approach: Query Models

- discriminative query models
  - learn robust support vector machine (SVM) classifier models
    - distinguishes query from rest/most of (subsampled) historic database
    - exploits invariance: “positive examples” include not only original query but also many shifted, scaled, time-warped versions of query
    - finds natural, query-relevant notions of similarity
      - focuses on ways query is unique from most historic data
      - also indicates feature weighting which would improve Euclidian distance-based similarity scores (for use in approx-NN indexing methods)
      - generative / probabilistic query models
  - learn state-transition models (e.g. HMM) of query behavior
  - facilitates handling missing or noisy sensor data
- hybrid approach: combine strengths of each
  - e.g. include match results from both; suggest features & variances to consider in other model types as well, …

Data and NASA Relevance

- we initially focus on two large time-series data sets:
  - IPEX-II space interferometer (SIM) boom structure
    - data obtained from: Dr. Marie Levine, JPL (Shuttle STS-85 payload)
    - initial set: 5 minutes of 1KHz for 24 accelerometer sensors (200,000 time points); total set: 10 Gbytes
    - relevance: IPEX thermal snap events represents case of rare phenomena to be harvested and understood from large data sets



  - Space Shuttle mission data
    - data obtained from: JSC (MEWS Shuttle data system)
    - currently working with hundreds of sensors from 3 recent missions
      - temperature and electrical sensors for STS 105,106, and 108
    - relevance: prime example of large-scale time-series NASA data set, with earlier-mission data similar to most latest-mission data

Accomplishments & Preliminary Findings

- key technical innovations to date:
  - efficient methods for training invariant-query SVMs models
    - enables many invariances and historic subsamples at query-time
  - radical speed-up (10-100x) of SVM classification times
    - enables complex query models that best discriminate between large numbers of query variants versus massive historic databases.
  - developed efficient methods for lazy generation of example vectors, from large space of rich time-series features
    - enables efficient batch and online training over large data sets, regardless of available computer RAM
    - enables feature selection over vast candidate spaces (for improved accuracy of query models and relevance of query matches)

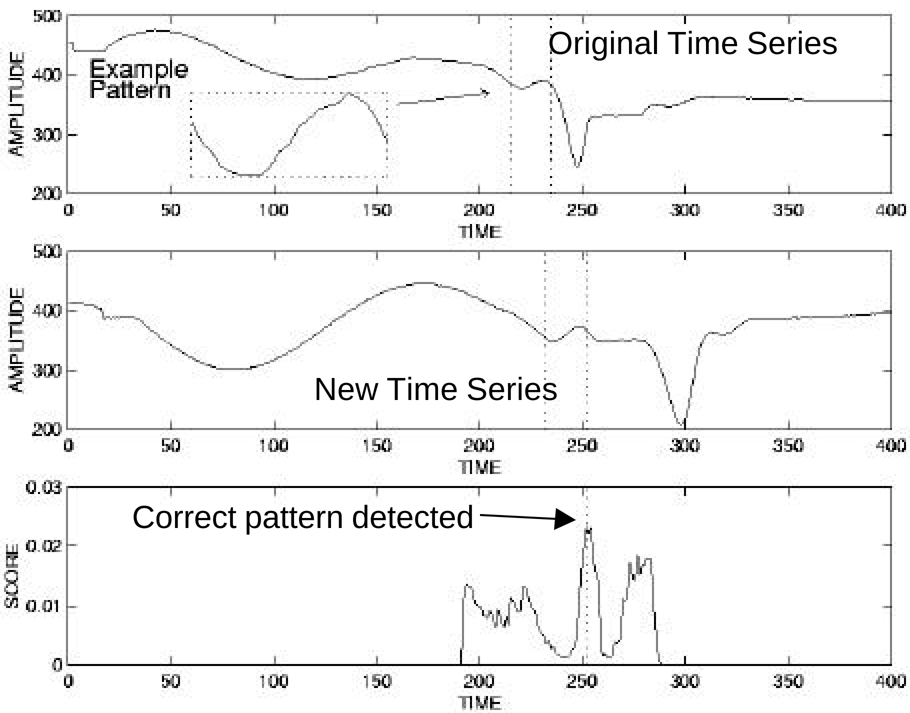
Accomplishments & Preliminary Findings (cont)

- Also, several advances on generative query models:
  - prior work (Ge and Smyth, ACM SIGKDD 2000):
    - probabilistic time-series query matching using probabilistic models

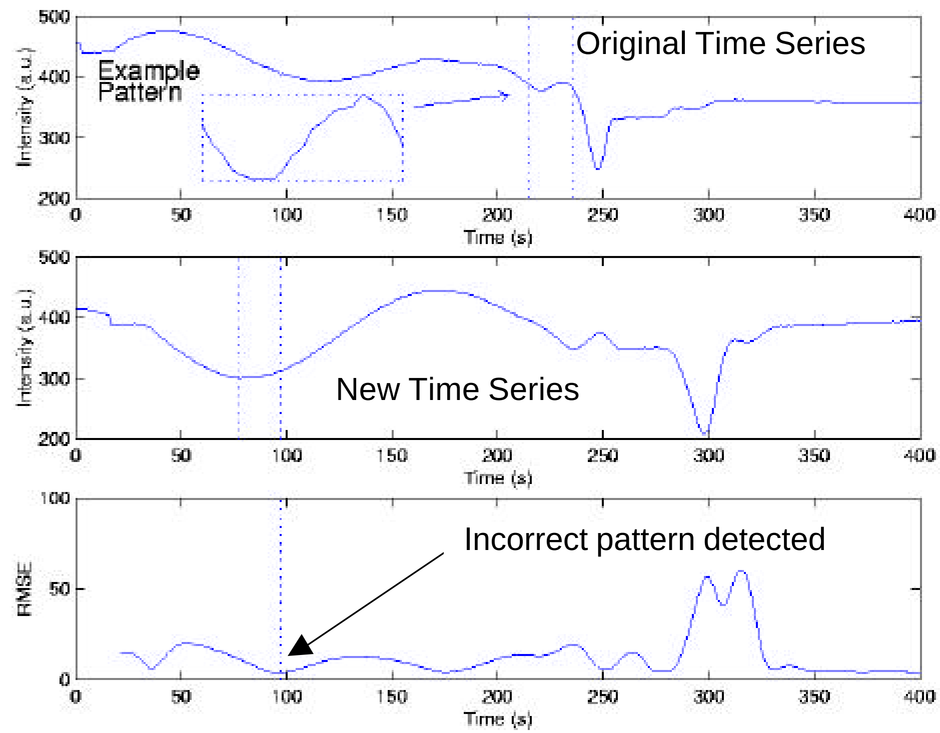
Accomplishments & Preliminary Findings (cont)

- example of benefit of generative query model vs traditional template matching

Detection Results with Markov Method



Detection Results with Template Matching



Technical Significance of Progress / Expected Impact on NASA

- our discriminative and generative query models provide query-relevant notions of similarity that capture user intention and task relevance much better than traditional *pre-query* measures of similarity.
- our SVM innovations, e.g. [Decoste, ICML-2002] giving orders of magnitude speedup of classification, are likely to have wide-spread impact on both the machine learning field and NASA.
- impact 1: makes SVMs competitive/superior speed-wise with popular alternatives (e.g. neural networks) for which SVMs have already been demonstrated to often be superior otherwise (accuracy, robustness).
- impact 2: makes SVMs practical in new applications (e.g. real-time classification onboard resource-constrained spacecraft

URLS Describing Team

- PI’s publications page:
  - http://www-aig.jpl.nasa.gov/home/decoste/dmd-pubs.html
- Co-I’s research group page:
  - http://www.datalab.uci.edu/

Facilities Used / Personnel

- JPL
  - Dr. Dennis DeCoste, PI
  - Dominic Mazzoni, computer scientist
  - 100-node Linux Beowulf machine, for testing parallel algos.
- University of California at Irvine:
  - Dr. Padhraic Smyth, co-I
  - Dasha Chudova, graduate student
  - Xianping Ge, graduate student

References

- Papers
  - D. DeCoste. **Anytime Interval-Valued Outputs for Kernel Machines: Fast Support Vector Machine Classification via Distance Geometry**. *Proceedings International Conference on Machine Learning* (ICML-02), July 2002.
  - D. DeCoste and B. Schoelkopf. **Training invariant support vector machines**, *Machine Learning* Journal, Volume 46(1-3), 2002.
  - D. DeCoste. **Visualizing Mercer kernel feature spaces via kernelized locally-linear embeddings**. The 8th International Conference on Neural Information Processing (ICONIP-2001). November 2001.
- Presentations
  - Invited tutorial, **Support Vector Machines and Other Kernel Methods: Key Concepts, Recent Advances, and Applications**, Institute for Pure and Applied Mathematics (IPAM), Conference on Mathematical Challeges in Scientific Data Mining, UCLA, January 17, 2002.